МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ РЫНОЧНОЙ СТОИМОСТИ НЕДВИЖИМОСТИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Максимовой Варвары Сергеевны

Научный руководитель:
старший преподаватель _____ Казачкова А.А.

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М. В.

подпись, дата

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. Актуальность темы подчеркивается необходимостью создания более точных и доступных решений для всех участников рынка недвижимости. В условиях глобализации и интеграции экономик пользователям требуется доступ к актуальным данным и инструментам для принятия обоснованных решений. Увеличение интереса к инвестициям в недвижимость создает потребность в глубоком анализе факторов, влияющих на стоимость объектов. Это открывает новые возможности для применения современных методов обработки данных, таких как машинное обучение, которое может значительно улучшить точность прогнозирования цен. Разработка эффективных алгоритмов для анализа и прогнозирования цен на недвижимость становится не только актуальной задачей, но и важным вкладом в развитие рынка недвижимости [1],[2].

Цель бакалаврской работы заключается в разработке приложения для прогнозирования рыночной стоимости недвижимости, которое основано на анализе влияния различных факторов на цену с использованием модели машинного обучения.

Поставленная цель определила **следующие задачи**:

1. Обзор существующих технологий прогнозирования рыночной стоимости недвижимости.
2. Обзор методов машинного обучения для прогнозирования цен.
3. Исследовательский анализ данных о недвижимости.
4. Использование регрессионных моделей для предсказания цен на жильё.
5. Применение ансамблевых методов, таких как случайный лес и градиентный бустинг, для улучшения точности предсказаний.
6. Создание веб-приложения с калькулятором цен на недвижимость для демонстрации результатов.

Методологические основы разработки приложения для прогнозирования рыночной стоимости недвижимости методами машинного обучения представлены в работах Naz R. (анализ эффективности алгоритмов

машинного обучения), Kong J. (применение нейронных сетей для прогнозирования цен), Tandon R. (сравнительный анализ регрессионных моделей), а также в исследованиях российских разработчиков коммерческих решений (Домклик, Авито), описанных в источниках [4, 5]. Теоретическая база по обработке данных и ансамблевым методам основана на трудах Маккинни У. (библиотека pandas), Бринк Х. (верификация гипотез), Элбон К. (полиномиальная регрессия) и Мюллер А**.** (реализация RandomForest, XGBoost).

Практическая значимость бакалаврской работы подтверждается разработкой веб-приложения с калькулятором стоимости недвижимости, интегрирующего модель XGBoost. Решение развернуто в Yandex Cloud с использованием Docker, обеспечивая масштабируемость и доступность для пользователей.

Структура и объём работы. Бакалаврская работа состоит из введения, 2 разделов (теоретический и практический), заключения, списка использованных источников и трех приложений. Общий объем работы – 120 страниц, из них 60 страниц – основное содержание, включая 28 рисунков и 3 таблицы, список использованных источников информации – 21 наименование.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические основы анализа данных и методов машинного обучения» посвящен комплексному исследованию современных подходов к прогнозированию цен на недвижимость, основанному на анализе актуальных научных работ и рыночных решений. В рамках исследования современных подходов к прогнозированию цен на недвижимость проведен детальный анализ актуальных научных работ и рыночных решений. Установлено, что регрессионные модели, в особенности линейная и полиномиальная регрессии, служат фундаментальным инструментом для выявления зависимостей между стоимостью объектов и их параметрами. При этом линейная регрессия позволяет количественно оценить влияние таких факторов, как площадь, этажность, тип дома и инфраструктурные характеристики. Однако критически важным аспектом является проблема мультиколлинеарности – ситуации, когда независимые переменные сильно коррелируют между собой, что может искажать интерпретацию коэффициентов модели. Полиномиальная регрессия, являясь частным случаем множественной линейной, расширяет возможности моделирования за счет учета нелинейных связей, хотя и усложняет интерпретацию отдельных параметров из-за высокой корреляции мономов.

Ансамблевые методы, включая Random Forest, Gradient Boosting и XGBoost, демонстрируют принципиально иные подходы к повышению точности прогнозов. Random Forest, основанный на бэггинге (bootstrap aggregating), создает множество деревьев решений, обученных на случайных подвыборках данных и признаков. Это обеспечивает устойчивость к переобучению и снижает дисперсию ошибок за счет агрегации предсказаний. Gradient Boosting, в свою очередь, использует последовательное обучение слабых моделей (обычно деревьев малой глубины) с фокусом на коррекции остаточных ошибок предыдущих итераций. XGBoost, как оптимизированная реализация градиентного бустинга, вводит регуляризацию (L1/L2) и эффективную обработку пропущенных значений, что значительно улучшает

производительность при работе с гетерогенными данными риелторских платформ. Ключевое преимущество ансамблей – способность комбинировать сильные стороны базовых алгоритмов: Random Forest минимизирует риск переобучения через рандомизацию, а бустинговые методы (особенно XGBoost) достигают высокой точности за счет адаптивной оптимизации функции потерь.

Система оценки качества регрессионных моделей включает комплекс метрик, каждая из которых акцентирует определенные аспекты ошибок. Среднеквадратичная ошибка (MSE) усиливает влияние крупных отклонений за счет квадратичной зависимости, что критично для задач, где недопустимы значительные ошибки в оценке дорогостоящих объектов. Корень из MSE (RMSE) сохраняет эту чувствительность, но выражается в единицах измерения исходных данных, упрощая интерпретацию. Средняя абсолютная ошибка (MAE), в отличие от них, линейно штрафует отклонения и менее чувствительна к выбросам, что предпочтительно в финансовых расчетах, где ошибка в 10 млн рублей считается ровно вдвое более значимой, чем в 5 млн. Коэффициент детерминации R^2 измеряет долю дисперсии целевой переменной, объясненную моделью, и служит универсальным индикатором адекватности: значения выше 0.8 указывают на высокое соответствие модели данным, тогда как отрицательные значения сигнализируют о неприменимости подхода. Практический пример с прогнозированием цен 10 квартир наглядно демонстрирует, как сочетание метрик (MAE=49 243 руб., R^2=0.85) позволяет комплексно оценить эффективность модели.

Веб-технологии, такие как микрофреймворк Flask, предоставляют оптимальную основу для разработки легковесных приложений машинного обучения. Flask, в отличие от полноценных фреймворков вроде Django, предлагает минималистичную архитектуру, где разработчик самостоятельно выбирает компоненты (ORM, аутентификацию), что обеспечивает гибкость при интеграции ML-моделей. Его модульная структура поддерживает RESTful-интерфейсы, упрощает создание API и совместима с шаблонизатором

Jinja2 для динамической генерации HTML. Однако отсутствие встроенных инструментов безопасности (кроме базовой защиты от XSS) требует дополнительных усилий при реализации аутентификации пользователей. Serverless-архитектура (FaaS), напротив, переносит инфраструктурные задачи на облачных провайдеров: функции выполняются по триггерам (HTTP-запросы), что устраняет необходимость управления серверами. Ключевые концепции – "холодный старт" (задержка при инициализации неактивных функций), лимиты параллелизма и таймауты исполнения – определяют сценарии применения: подход идеален для обработки событий с переменной нагрузкой, но менее пригоден для долгих фоновых процессов.

Второй раздел «Практическая часть» включает исследовательский анализ данных, обучение модели и разработку приложения. Исследовательский анализ данных (EDA) начался с объединения информации из 20 городов России, включая Москву, Санкт-Петербург и Саратов, в единый датасет объемом 55 601 объявление. Первичная обработка выявила структурные проблемы: 56 столбцов (из 108) содержали более 50% пропусков, включая критичные атрибуты типа "ГодПостр" (39.8%) и "ТипКомнат" (48.9%). Принято решение об удалении этих признаков, поскольку импутация медианой или модой могла внести существенные смещения. Дубликаты (1.05% записей) идентифицированы по текстовым полям "Заголовок" и "Описание" и удалены с сохранением последней версии объявления. Глубокий анализ распределений выявил ключевые закономерности рынка: доминирование 1–3-комнатных квартир (68.6% выборки), преобладание кирпичных домов (37.7%) и выраженную асимметрию ценовых показателей. Визуализация через тепловые карты корреляций и боксплоты подтвердила, что стоимость объектов значимо зависит от типа дома (монолитные здания дороже панельных на 15–20%), наличия балкона/лоджии и этажа – например, в кирпичных домах цена растет на верхних этажах, тогда как в панельных наблюдается обратная тенденция.

Географический анализ обнаружил концентрацию предложений в крупных городах: Москва и Санкт-Петербург совокупно формировали свыше 25% данных. Сравнение цен по регионам выявило двукратную разницу: медианная стоимость в Москве (15.2 млн руб.) существенно превышала показатели Хабаровска (7.8 млн) или Калининграда (8.1 млн). При этом обнаружена обратная зависимость между количеством объявлений и средней ценой – города с насыщенным рынком (Хабаровск) демонстрировали более низкие значения, что указывает на конкурентное ценообразование. Пространственная визуализация через интерактивные карты (библиотека Plotly) выявила кластеры дорогой недвижимости в центрах мегаполисов, что подтверждает гипотезу о значимости локации.

Прогнозирование стоимости реализовано в несколько этапов. Для балансировки данных по малым городам применена синтетическая аугментация: алгоритм генерировал новые записи на основе существующих объектов, варьируя цену (±15%), площадь (±10%) и категориальные признаки (тип дома, ремонт). Предобработка включала удаление выбросов методом межквартильного размаха (IQR), что исключило 7.4% аномальных записей, заполнение пропусков в числовых полях медианами и создание бинарных признаков ("Последний этаж", "с/х местность"). Категориальные переменные закодированы via One-Hot Encoding, а числовые – стандартизированы через StandardScaler. Сравнение моделей показало, что композитный подход (объединение предсказаний KNN по координатам с основными признаками) уступает чистому XGBoost (R^2=0.64). Дальнейшая оптимизация гиперпараметров методом рандомизированного поиска (RandomizedSearchCV) с настройкой количества деревьев, глубины и скорости обучения улучшила метрики до R^2=0.86 и MAE=662 705 руб. Анализ ошибок выявил регионы с максимальными отклонениями (Заречье – MAE 8.1 млн руб.), что связано с уникальностью объектов premium-сегмента. Отбор признаков по важности сократил их число до ключевых (геокоординаты, тип ремонта, этажность), сохранив R^2=0.61 при упрощении интеграции.

Веб-приложение разработано на базе Flask с трехслойной архитектурой:

1. **Клиентский интерфейс** (HTML/CSS/JS) – включает интерактивную форму с валидацией, выбором адреса через Яндекс.Карты и визуализацией типов ремонта через изображения.
2. **Серверный слой** (Python/Flask) – обрабатывает запросы, интегрируется с API Яндекс.Геокодера для преобразования адреса в координаты и использует предобученную модель XGBoost.

Ключевые функции: история оценок с сохранением в cookies, анимация результатов, адаптивный дизайн. Развертывание автоматизировано через Docker-контейнеризацию: образ включает зависимости Python, модель и веб-интерфейс. Для публикации выбран Yandex Cloud в режиме Serverless Containers, что исключило необходимость управления серверами и обеспечило масштабируемость под нагрузкой. Решение демонстрирует эффективность связки "микрофреймворк + контейнеризация + облачные сервисы" для задач ML в реальном времени.

ЗАКЛЮЧЕНИЕ

В ходе ВКР были последовательно выполнены все поставленные задачи. Проведен анализ современных исследований и существующих решений, что позволило определить ключевые подходы к прогнозированию цен. Исследовательский анализ данных выявил основные закономерности рынка, включая зависимость стоимости от площади, местоположения, типа дома и инфраструктурных характеристик. Сравнительный анализ цен в разных регионах подтвердил значимость географического фактора, а применение регрессионных и ансамблевых моделей (линейная регрессия, градиентный бустинг, XGBoost) продемонстрировало их эффективность. Наилучшие результаты показала модель XGBoost, достигнутая точность которой (R^2 = 0.61) подтверждает ее применимость для решения практических задач. Разработанное веб-приложение на базе Flask интегрирует полученную модель, предоставляя пользователям удобный инструмент для оценки стоимости недвижимости в режиме реального времени.

Цель работы — создание приложения для прогнозирования цен на недвижимость — достигнута. Реализованное решение учитывает разнородные факторы, включая технические параметры объектов, географические данные и рыночные тренды. Использование Docker и облачных технологий (Yandex Cloud) обеспечило масштабируемость и отказоустойчивость системы.

Перспективы дальнейшего развития работы связаны с расширением функционала приложения за счет интеграции дополнительных данных (инфраструктура района, экологические показатели) и применения методов глубокого обучения для анализа текстовых описаний объектов. Результаты исследования вносят вклад в развитие методов анализа данных в сфере недвижимости, подтверждая потенциал машинного обучения для решения сложных экономических задач.

Основные источники информации:

1. Naz, R. Real Estate Price Prediction / R. Naz // Applied and Computational Engineering. — 2024. — Vol. 6. — P. 1031–1044.

2. Kong, J. House price prediction / J. Kong // Applied and Computational Engineering. — 2024. — Vol. 75. — P. 141–146. DOI: 10.54254/2755-2721/75/20240526.

3. Tandon, R. The Machine Learning Based Regression Models Analysis For House Price Prediction / R. Tandon // International Journal on Document Analysis and Recognition (IJDAR). — 2024. — Vol. 11. — P. 296.

4. Определение рыночной цены квартиры [Электронный ресурс]: [сайт]. — URL: https://price.domclick.ru (дата обращения: 05.11.2024). — Загл. с экрана. — Яз. рус.

5. Определение рыночной стоимости квартиры [Электронный ресурс]: [сайт]. — URL: https://www.avito.ru/evaluation/realty/ (дата обращения: 05.11.2024). — Загл. с экрана. — Яз. рус.

6. Маккинни, У. Python и анализ данных. Первичная обработка данных с применением pandas, NumPy и Jupyter; пер. А. А. Слинкин. — 3-е изд. — Москва : ДМК Пресс, 2023. — 537 с. — Загл. с экрана. — Яз. рус.

7. Бринк, Х. Машинное обучение / Х. Бринк, Дж. Ричардс, М. Феверолф. — СПб. : Питер, 2017. — 480 с.